Estimating occupancy dynamics and encounter rates with species misclassification: a semi-supervised individual-level approach

Anna I. Spiers [a,b], J. Andrew Royle [c], Christa L. Torrens [d], Maxwell B. Joseph [a]

[a] Earth Lab, Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, CO, USA

[b] Department of Ecology and Evolutionary Biology, University of Colorado Boulder, CO, USA

[c] US Geological Survey Patuxent Wildlife Research Center

[d] Institute of Arctic and Alpine Research, University of Colorado, Boulder, CO, USA

Correspondence

Anna Spiers

Email: anna.spiers@colorado.edu

Running headline

Species classification-occupancy model

Abstract

1. Large-scale, long-term biodiversity monitoring is essential to meeting conservation and land management goals and identifying threats to biodiversity. However, multispecies surveys are prone to various types of observation error, including false positive/negative detection, and misclassification, where a species is encountered but its species identity is not correctly identified. Previous methods assume an imperfect classifier produces species-level classifications, but in practice, particularly with human observers, we may end up with extraspecific classifications including "unknown", morphospecies designations, and taxonomic identifications coarser than species. Disregarding these types of species misclassification in biodiversity monitoring datasets can bias estimates of ecologically important quantities such as detectability, occurrence, and species richness.
2. Here we develop an occupancy model that accounts for species non-detection and misclassification. Our framework accommodates extinction and colonization dynamics, allows for additional uncertain "morphospecies" designations in the imperfect species classifications, and makes use of individual specimen with known species identities in a semi-supervised setting. We assess the performance of our joint classification-occupancy model to a reduced classification model that discards information about occupancy and encounter rate on a withheld test set. We illustrate our model with an empirical case study of the carabid beetle community at the NEON Niwot Ridge site and quantify taxonomist identification error by accounting for classification probabilities.
3. Species occupancy varied through time and across sites and species. The model yielded high probabilities of classification where the imperfect classifier matched the true species. The classification model informed by occupancy and encounter rates outperformed the classification that was not, and these differences were most pronounced for abundant species.
4. Our probabilistic framework can be applied to datasets with imperfect species detection and classification. This model can identify commonly misclassified species, helping biodiversity monitoring organizations systematically prioritize which samples need validation by an expert. Our Bayesian approach propagates classification uncertainty to offer an alternative to making conservation decisions based on point estimates

Keywords — carabid, imperfect classifier, morphospecies, NEON, observation error, occupancy models, semi-supervised, species misclassification

1 Introduction

Large-scale, long-term biodiversity monitoring is essential to meeting conservation and land management goals and identifying threats to biodiversity. Such comprehensive datasets increasingly include multispecies surveys that capture information-rich co-occurrence data, enabling community-level analyses (Iknayan et al., 2014; Ovaskainen et al., 2017).

However, multispecies surveys are prone to various types of imperfect detection, including false absences where a species is present but not detected (Dorazio and Royle, 2005), and misidentification, where a species is encountered but its species identity is not correctly recorded (Miller et al., 2011).

Occupancy models account for observation error in biodiversity surveys that seek to understand species distributions, track population changes, and describe mechanisms underlying population and community dynamics (MacKenzie et al., 2002). Latent presence/absence states are modeled explicitly, with an observation model that accounts for the details of the detection process, including the potential for false negatives (non-detections at occupied sites) and false positives (detections at unoccupied sites) (Royle and Link, 2006; Miller et al., 2012; Chambert et al., 2015; Wright et al., 2020). Disregarding false positives in biodiversity monitoring datasets can bias estimates of ecologically important quantities such as detectability, occurrence, and species richness (McClintock et al., 2010; Chambert et al., 2015, 2018).

Multi-species surveys are also subject to errors in species identifications by imperfect classifiers. Imperfect classifiers include citizen scientists (e.g., North American Breeding Bird Survey (Sauer et al., 2017)), technicians trained in local taxonomy (e.g., invertebrate trapping by NEON (Hoekman et al., 2017)), automated methods (e.g., bat acoustic recording software (Wright et al., 2020) or convolutional neural networks used with camera trap data (Tabak et al., 2019)). Previous methods assume an imperfect classifier produces species-level classifications, but in practice, particularly with human observers, we may end up with extraspecific classifications including "unknown", morphospecies designations, and taxonomic identifications coarser than species.

If species are prone to misclassification, then known species identities might be used to correct estimates of occupancy parameters. However, using these data presents a methodological challenge. We refer to this situation as "semi-supervised": true species identities are known for some but not all individuals. Previous multi-species occupancy models that accommodate misclassification have used multinomial models that sum over all individuals (Wright et al., 2020), or site-level validation data where the occupancy state of a species is known only at a site- or plot-level but not at an individual-level (Chambert et al., 2018). Using individual-level validation data requires a different approach.

Misclassified species identities can be dealt with using one of two contrasting approaches. A simple two step approach 1) uses a classifier to assign species IDs to each individual (creating one complete synthetic dataset from classifier output, for which species identities are treated as known), then 2) analyzes the constructed dataset using a downstream model (e.g., an occupancy model). This two step approach does not propagate uncertainty in species identity to the downstream model, and the assignment of species identities in the first stage does not use any information about occupancy or encounter rates. In contrast, a joint model directly uses classifier output as data, relating the observation process to underlying ecological states in one step. Such an approach can simultaneously account for uncertainty in species identities, and leverage information about occupancy and encounter rates to inform species identity estimates (Wright et al., 2020). However, there remains the practical question of how much value is added by a joint model vs. a two-stage approach. *A priori*, we expect that a joint model should produce better

estimates of true species identities by using information on occupancy and encounter rates, but this has not yet been tested.

Here we develop an individual-level, semi-supervised, dynamic occupancy model that accounts for species non-detection and misclassification. Our Bayesian approach propagates classification uncertainty to offer an alternative to making conservation decisions based on point estimates. Our framework extends the classification occupancy model of Wright et al. (2020) to 1) accommodate extinction and colonization dynamics, 2) allow for additional uncertain "morphospecies" designations in the imperfect species classifications, and 3) make use of labeled samples with known species identities in a semi-supervised setting. Further, we compare the performance of a classification occupancy model to a reduced classification model that discards information about occupancy and encounter rate on a withheld test set. We demonstrate our model with an empirical case study of the carabid beetle (*Carabidae*) community at the National Ecological Observatory Network (NEON) Niwot Ridge site and quantify taxonomist identification error by accounting for classification probabilities.

2 Materials and Methods

2.1 Modeling occupancy dynamics with misclassification

Consider data collected at sites $i = 1, ..., N$, according to a robust design (Hoekman et al., 2017) where each site is visited J times within primary periods $t = 1, ..., T$, where the occupancy states are assumed to be constant within primary periods.

2.1.1 State model

We are interested in occupancy states and encounter rates for species $k = 1, ..., K$. Sites are either occupied ($z_{i,k,t} = 1$) or not ($z_{i,k,t} = 0$). We assume that the occupancy states arise as Bernoulli random variables:

$$z_{i,k,t} \sim \text{Bernoulli}(\psi_{i,k,t}).$$

The probability of occupancy in the initial primary period is $\psi_{i,k,1}$. Subsequent occupancy dynamics depend on the probability of colonization $\gamma_{i,k,t}$ and persistence $\phi_{i,k,t}$, such that for $t > 1$:

$$\psi_{i,k,t} = z_{i,k,t-1}\phi_{i,k,t-1} + (1 - z_{i,k,t-1})\gamma_{i,k,t-1}$$

2.1.2 Encounter model

On any particular sampling occasion j, we encounter $L_{i,j,k,t}$ individuals with encounter rate $\lambda_{i,j,k,t}$. We assume that the number of encounters is a Poisson random variable: $L_{i,j,k,t} \sim \text{Poisson}(z_{i,k,t}\lambda_{i,j,k,t})$. In a setting with misclassification, the number of encountered individuals $L_{i,j,k,t}$ is not observed directly because of uncertainty in the true species identities of encountered individuals. We do however observe the total number of individuals encountered

on any particular occasion: $L_{i,j,.,t} = \sum_{k=1}^{K} L_{i,j,k,t}$. The properties of sums of Poisson random variables allow us to model these observed totals as:

$$L_{i,j,.,t} \sim \text{Poisson}\Big(\sum_{k=1}^{K} z_{i,k,t}\lambda_{i,j,k,t}\Big).$$

2.1.3 Observation model

In addition to observing the total number of encountered individuals on an occasion $L_{i,j,.,t}$, we assume that we also obtain imperfect species classifications for each encountered individual. In cases where individuals have been encountered ($L_{i,j,.,t} > 0$), we obtain imperfect classifications of individuals $l = 1, ..., L_{i,j,.,t}$ and model these as arising from a categorical distribution with a species-specific probability vector:

$$y_{i,j,l,t} \sim \text{Categorical}(\boldsymbol{\theta}_{k[i,j,l,t]}),$$

where $y_{i,j,l,t}$ is the imperfect classification, and $\boldsymbol{\theta}_{k[i,j,l,t]}$ is a probability vector associated with the true species of individual l, which we denote $k[i, j, l, t]$. Element k' in the vector $\boldsymbol{\theta}_{k[i,j,l,t]}$ represents the probability that an individual is classified into category k', conditional on the true species identity $k[i, j, l, t]$, such that $\theta_{k[i,j,l,t],k'} = \Pr(y_{i,j,l,t} = k' \mid k[i, j, l, t])$. If species are always misclassified as other species, then θ_k will be a vector of length K (Wright et al., 2020). If there are extraspecific classes (e.g. morphospecies), θ_k may have more than K elements.

True species identities are modeled as:

$$k[i, j, l, t] \sim \text{Categorical}\left(\frac{z_{i,k,t}\lambda_{i,j,k,t}}{\sum_k z_{i,k,t}\lambda_{i,j,k,t}}\right).$$

If ground truth species identity data are available for some individuals, then $k[i, j, l, t]$ is partly observed and this model can be trained in a semi-supervised setting. In the unsupervised setting, this individual-level formulation is a disaggregated version of the single-season multinomial model of Wright et al. (2020) (Appendix S1).

2.1.4 Incorporating morphospecies designations

In some settings the imperfect classifier might assign more classes than there are unique species, so that the vector $\boldsymbol{\theta}_k$ has more than K elements. For example, in the NEON beetle data, if a parataxonomist is unable to identify a set of similar individuals, they will classify those individuals as a unique morphospecies associated with that sampling occasion. Thus, it is possible for individuals to be classified into $\tilde{K} \geq K$ classes, where $\tilde{K}$ is sum of the number of species and the total number of morphospecies designations. In such cases, the matrix $\boldsymbol{\Theta} = (\boldsymbol{\theta}'_1, ..., \boldsymbol{\theta}'_K)$ can be rectangular, with the first K columns corresponding to the classification probabilities for species $1, ..., K$, and the remaining columns corresponding to classification probabilities for non-species classes, e.g., morphospecies:

$$\boldsymbol{\Theta} = \overbrace{\begin{bmatrix} \theta_{1,1} & \dots & \theta_{1,K} & \dots & \theta_{1,K'} \\ \vdots & \ddots & & \ddots & \\ \theta_{K,1} & & \theta_{K,K} & & \theta_{K,K'} \end{bmatrix}}^{\substack{\text{Species} \\ \text{classifications}}} \overbrace{\phantom{\begin{bmatrix} \theta_{1,1} & \dots & \theta_{1,K} & \dots & \theta_{1,K'} \\ \vdots & \ddots & & \ddots & \\ \theta_{K,1} & & \theta_{K,K} & & \theta_{K,K'} \end{bmatrix}}}^{\substack{\text{Morphospecies} \\ \text{classifications}}} \Bigg\} \begin{array}{l} \text{True} \\ \text{species} \\ \text{identities} \end{array}$$

2.2 Case study

2.2.1 Application to NEON carabid data

We fit our model to the carabid pitfall trap sampling data collected by NEON at Niwot Ridge in CO, USA in 2015-2018 (NEON, 2021). Niwot Ridge is a site in the southern Rocky Mountains, spanning subalpine conifer forest and alpine tundra. We outline the relevant data collection protocol here, but Hoekman et al. (2017) offer more detail regarding NEON's carabid pitfall trap data product. The sampling design at every NEON site consists of ten permanent plots across the site with four pitfall traps per plot. Traps are sampled and reset biweekly during the growing season, with a range of 5-7 collections per year at Niwot Ridge. Our model runs at the plot-level. In 2018 one plot was permanently relocated to ensure sampling was allocated proportionally to the NLCD cover types represented (NEON help desk, personal communication).

All carabid samples are classified by a parataxonomist, and a subset are sent to an expert taxonomist for validation (Figure 1) (Hoekman et al., 2017). Species classification by parataxonomists is considered imperfect due to the brief taxonomic training of parataxonomists. Identification by an expert taxonomist is treated as confirmation data and is limited due to budget constraints. We confirmed the accuracy of the expert taxonomist classifications in finding that all individuals sent for DNA barcoding by NEON match the expert taxonomist's identification for the samples we used. In the few cases where the expert taxonomist could not identify a specimen to species-level, we use their genus-level classification for the validation dataset.

Our dataset contains 4772 individuals, 1764 of which were identified by an expert taxonomist, and 62 species classified by the parataxonomist, 23 of which are morphospecies. Morphospecies identifications are unique to each NEON site and year. We fit our model using all individuals and used no environmental covariates. A hurdle for the NEON community in using the carabid pitfall trap data is reconciling parataxonomist and expert taxonomist classifications (Figure 1). Only one study to date has been published using the NEON carabid pitfall trap data (Egli et al., 2020), but Egli et al. (2020) analyze only the subset of individuals that have expert taxonomist classifications. Our model lowers the barrier of entry for using all imperfectly classified individuals while leveraging the available validation data.

2.2.2 Model specification

We used informative priors for the species classification probability vectors $\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K$ that placed higher probability density on the correct species classification. In the case of NEON beetle data, this is reasonable given the training

that parataxonomists receive in beetle identification. Because all elements of each $\boldsymbol{\theta}_k$ vector need to sum to one, and each element is bounded between 0 and 1, we used a Dirichlet prior: $\boldsymbol{\theta}_k \sim \text{Dirichlet}(\boldsymbol{\alpha}_k)$. We chose the Dirichlet concentration values $\boldsymbol{\alpha}_k$ by comparing draws from the Dirichlet prior distribution to our prior intuition about parataxonomist skill, using 200 along the diagonal (the element corresponding to the correct species identity), and 2 elsewhere.

We used multivariate normal priors at the species and site level, which allowed for correlations among parameters. These priors share information among initial occupancy, persistence, colonization, and encounter rates. The motivation for this stemmed from a prior expectation that these parameters could be related. For example, species that are more abundant might be more likely to occur initially, persist, or colonize new sites. Similar arguments could be made about relationships among parameters at a site level.

Each species is associated with a vector $\boldsymbol{\alpha}_k$ of length 4, where $\alpha_{k,1}$, $\alpha_{k,2}$, $\alpha_{k,3}$, and $\alpha_{k,4}$ are species-specific adjustments on initial occupancy, persistence, colonization, and encounter rates respectively. We assume that the species-specific adjustments are drawn from a multivariate normal prior with mean equal to zero, and an unknown covariance matrix: $\boldsymbol{\alpha}_k \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}^{(\alpha)})$. Similarly, site-specific adjustments $\boldsymbol{\epsilon}_i$ were drawn from a different multivariate normal prior. These adjustments were added together on a transformed scale to compute initial occupancy, persistence, colonization, and encounter rates, e.g., $\text{logit}(\psi_{i,k,1}) = \epsilon_{i,1} + \alpha_{k,1}$. A full model specification for the case study is available in Appendix S2 (Plummer et al., 2003).

To evaluate how the occupancy and encounter rate components of the full model informed classification probability estimates, we also developed a reduced model that discards all information about occupancy and abundance, using just the expert and para-taxonomist species classifications to estimate the classification matrix $\boldsymbol{\Theta}$. This comparison reveals the extent to which occupancy and encounter rates inform classification probabilities. If there are no differences in the estimates of classification probabilities, then a two-stage model which first models misclassification and then passes the posterior on as a prior for an occupancy/encounter model should perform as well as the joint model in which the classification model is integrated with the occupancy model. In addition to comparing posterior estimates for $\boldsymbol{\Theta}$, we withhold a randomly selected 20% of the imperfect classifications to evaluate which model (full or reduced) better predicts the data generated by the parataxonomist. All models were fit using JAGS, dclone, and R v4.0.2 (Plummer et al., 2003; Sólymos, 2010; R Core Team, 2020) and visualized with ggplot (Wickham, 2016).

3 Results

3.1 Dynamic occupancy model

The occupancy model was designed to allow correlation between parameters across sites and species. occupancy, growth, and turnover rates also varied through time. Sites with high encounter rates tended to have low initial occupancy and colonization probabilities and high persistence probabilities (Figure 2). Further, sites with high colonization rates tended to have high initial occupancy probabilities and low persistence probabilities. At the species

level, we saw positive correlations among many of the model components, but in particular, species' encounter rate was positively correlated with species' initial occupancy, persistence, and colonization rates (Figure 2). Species varied from abundant and accurately identified by the parataxonomist (e.g. *Calathus advena*) to not once identified by the parataxonomist (e.g., *Dicheirotrichus mannerheimii*).

3.2 Classification model

The model yielded high probabilities of classification along the diagonal of the θ confusion matrix where the expert and para-taxonomist identifications match (Figure 3). The model favors the parataxonomist's skill by giving more weight in the theta prior to diagonal values, making morphospecies classifications less probable. Individuals with morphospecies classifications make up a sizeable portion of the community, 811 out of the total 4772 total individuals identified by the parataxonomist. Despite the dirichlet priors favoring parataxonomist accuracy, some species had nontrivial probabilities of being classified as morphospecies than as the true species by the parataxonomist. For example, the parataxonomist was more likely to classify *Pterostichus (Hypherpes) sp.* as D13.2016.MorphBT than as the true species (Figure 3). However, no species had more than 3% probability (median) of being classified as another species (i.e., our model results indicate that the parataxonomist is most likely to identify a species either correctly or as a morphospecies).

To evaluate the value added by informing the classification model with occupancy and encounter rates, we compared the full model to a reduced classification model that discards all information about occupancy and abundance. Most θ_k probability vectors do not differ between the full and reduced model results. However, we see differences for a few species where there are non-overlapping θ posterior density distributions between the full and reduced models (i.e., Theta[46,46] and Theta [48,48], Figure 4). These differences are found most notably for the abundant species. The full model yielded higher classification probabilities for the abundant species. Further, the reduced model has wider 95% credible intervals compared to the full model for many theta indices (Figure 5). Thus, we find that a joint occupancy-classification model outperforms a two-stage model (classification, then occupancy).

We evaluated the performance of the full and reduced models by withholding a randomly selected 20% (352 individuals) of the imperfect classifications that have an expert identification (1764 individuals). For the withheld individuals, the validation metric macro-averages are listed in Table 1. For every validation metric, the full model yields better results than the reduced model. The validation metrics calculated to the species level highlighted substantial differences between the models for common species. This confirms the result that occupancy dynamics improve classification model performance compared to using classification data alone.

4 Discussion

We developed a statistical approach to improve classification in multispecies datasets that leverages occupancy dynamics. Our probabilistic framework can be applied to datasets with imperfect species detection and also errors

in classification of samples. This approach builds on recent work on classification in occupancy models (Devarajan et al., 2020, and references therein) by evaluating the advantage of a joint occupancy-classification model, allowing imperfect classifications to outnumber species, and leveraging individual-level confirmation data in a semi-supervised setting. While analyses targeting species richness may be shielded to a certain extent from imperfect classification (Egli et al., 2020), any population- or community-level analyses with taxonomic specificity require an understanding of classification uncertainty in the data. Whereas imperfect classifiers offer classification point-estimates, our model provides a vector of probabilities for every species.

This is the first model to consider how occupancy and encounter rates contribute to improving classification. We found that our joint occupancy-classification model outperformed a reduced model that disregarded occupancy dynamics in estimating imperfect classification (Figure 6, Table 1). When looking at the validation metrics at the species level, the joint model surpasses the reduced model even more for abundant species. In Figure 5, we see that the full model posteriors have smaller CI widths than the reduced model for many species, which visualizes the superior precision of the full model.

False positive and negative species classifications are inevitable in any field collection, due to time and money constraints or imperfect classifiers (Royle and Link, 2006; Miller et al., 2012; McClintock et al., 2010; Hoekman et al., 2017). Accounting for false identifications is important to reduce bias in occupancy dynamics estimated from multispecies biodiversity monitoring datasets (McClintock et al., 2010; Miller et al., 2011; Chambert et al., 2015; Miller et al., 2015). Alternative models that account for false positives may consider data from only the focal species (Chambert et al., 2015) or from binary observations (Chambert et al., 2017). Like Wright et al. (2020), we use available counts from an imperfect classifier (Figure 1). However, we use all species detected, no matter how rare. By allowing taxonomic uncertainty propagation for multispecies datasets where imperfect classifications outnumber species (e.g., unknown, morphospecies, to the family level) by using a rectangular classification matrix θ (Figure 3), we remove an assumption that previous occupancy modeling methods have used (Chambert et al., 2018; Wright et al., 2020).

Our model is semi-supervised and runs on data at the individual-level. Whereas alternative models use data pooled at the site- or occasion-level (Chambert et al., 2018; Wright et al., 2020), our model leverages the rich information at the individual-level to reveal which species are commonly mistaken by the imperfect classifier and how often. Individuals identified by the expert taxonomist we know as true positives so can be used as partially-observed occupancy data in our semi-supervised model. In contrast, the model by Wright et al. (2020) is unsupervised, but also at the individual-level. Because our analysis is done at the individual-level, we can use species counts to inform classification (Chambert et al., 2017). Although the model priors favor parataxonomist accuracy, the model found high probability of classification for a couple of morphospecies that were abundant in the data (D13.2015.MorphO and D13.2016.MorphBT) (Figure 3).

Despite its innovations, our model has limitations. For NEON's ground beetle data from Niwot Ridge, which we used to fit our model, the parataxonomists were skilled and had high identification agreement with expert taxonomist

classifications. The model may yield unexpected results when applied to a NEON site with lower parataxonomist accuracy. This begs the question of how much validation data is necessary to fit the model for varying degrees of imperfect classifier accuracy. This could be answered with simulations to identify what percentage or which type of samples should be prioritized for expert taxonomist classification to yield desired results. We tried various iterations of the model before coming to the final disaggregated, semi-supervised, individual-level model. Using an aggregated data approach, we found the model either would not converge or struggled with identifiability, yielding multimodal posteriors for θ. Changing the Dirichlet priors to favor parataxonomist accuracy helped but did not eliminate the problem. Future work could more explicitly incorporate false positives by informing the θ priors with a list of species commonly misidentified by the imperfect classifier.

Large-scale, long-term biodiversity surveys are critical to inform land management and conservation policy (Hughes et al., 2017) and require affordable and efficient species classifications to stay on track and within budget. Egli et al. (2020) make a case for better training of parataxonomists to improve classification error rates. However, training people is expensive and staff can be transient, calling for a more systematic solution. This probabilistic approach can model species occupancy while accounting for imperfect classification, without additional training. Innovations in occupancy models in general, are rapidly being made to consider an expanding variety of study systems and experimental designs (Bailey et al., 2014). Our results support that ecological dynamics (i.e. occupancy and encounter rates) inform classification probabilities and lays a foundation for future work to build upon.

Acknowledgements

The National Ecological Observatory Network is a program sponsored by the National Science Foundation and operated under cooperative agreement by Battelle Memorial Institute. This material is based in part upon work supported by the National Science Foundation through the NEON Program. We thank G. Vagle for taking part in the conception of this project and J. Coulombe for her graphical design assistance. The work was supported by the CU Boulder Grand Challenge investment in Earth Lab. AIS was supported as a GRA at Earth Lab for work on this project. Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Authors' contributions

AIS, CLT, and MBJ conceived the project idea; AIS, JAR, and MBJ designed methodology; AIS curated the data; AIS and MBJ analysed the data; AIS and MBJ led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data availability

Carabid data accessible through the NEON data portal at https://data.neonscience.org/data-products/DP1.10022.001 (NEON, 2021). Code for data cleaning and analysis is available at https://github.com/annaspiers/NEON-NIWO-misclass

References

Bailey, L. L., MacKenzie, D. I., and Nichols, J. D. (2014). Advances and applications of occupancy models. *Methods in Ecology and Evolution*, 5(12):1269–1279. https://doi.org/10.1111/2041-210X.12100.

Chambert, T., Grant, E. H. C., Miller, D. A., Nichols, J. D., Mulder, K. P., and Brand, A. B. (2018). Two-species occupancy modelling accounting for species misidentification and non-detection. *Methods in Ecology and Evolution*, 9(6):1468–1477. https://doi.org/10.1111/2041-210X.12985.

Chambert, T., Miller, D. A., and Nichols, J. D. (2015). Modeling false positive detections in species occurrence data under different study designs. *Ecology*, 96(2):332–339. https://doi.org/10.1890/14-1507.1).

Chambert, T., Waddle, J. H., Miller, D. A., Walls, S. C., and Nichols, J. D. (2017). A new framework for analysing automated acoustic species detection data: Occupancy estimation and optimization of recordings post-processing. *Methods in Ecology and Evolution*, 9(3):560–570. https://doi.org/10.1111/2041-210X.12910.

Devarajan, K., Morelli, T. L., and Tenan, S. (2020). Multi-species occupancy models: review, roadmap, and recommendations. *Ecography*. https://doi.org/10.1111/ecog.04957.

Dorazio, R. M. and Royle, J. A. (2005). Estimating size and composition of biological communities by modeling the occurrence of species. *Journal of the American Statistical Association*, 100(470):389–398. https://doi.org/10.1198/016214505000000015.

Egli, L., LeVan, K. E., and Work, T. T. (2020). Taxonomic error rates affect interpretations of a national-scale ground beetle monitoring program at national ecological observatory network. *Ecosphere*, 11(4):e03035. https://doi.org/10.1002/ecs2.3035.

Hoekman, D., LeVan, K. E., Ball, G. E., Browne, R. A., Davidson, R. L., Erwin, T. L., Knisley, C. B., LaBonte, J. R., Lundgren, J., Maddison, D. R., et al. (2017). Design for ground beetle abundance and diversity sampling within the national ecological observatory network. *Ecosphere*, 8(4):e01744. .

Hughes, B. B., Beas-Luna, R., Barner, A. K., Brewitt, K., Brumbaugh, D. R., Cerny-Chipman, E. B., Close, S. L., Coblentz, K. E., De Nesnera, K. L., Drobnitch, S. T., et al. (2017). Long-term studies contribute disproportionately to ecology and policy. *BioScience*, 67(3):271–281. https://doi.org/10.1002/ecs2.1744.

Iknayan, K. J., Tingley, M. W., Furnas, B. J., and Beissinger, S. R. (2014). Detecting diversity: emerging methods to estimate species diversity. *Trends in ecology & evolution*, 29(2):97–106. https://doi.org/10.1016/j.tree.2013.10.012.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. (2002). Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 83(8):2248–2255. .

McClintock, B. T., Bailey, L. L., Pollock, K. H., and Simons, T. R. (2010). Unmodeled observation error induces bias when inferring patterns and dynamics of species occurrence via aural detections. *Ecology*, 91(8):2446–2454. https://doi.org/10.1890/0012-9658(2002)083[2248:ESORWD]2.0.CO;2.

Miller, D. A., Bailey, L. L., Grant, E. H. C., McClintock, B. T., Weir, L. A., and Simons, T. R. (2015). Performance of species occurrence estimators when basic assumptions are not met: a test using field data where true occupancy status is known. *Methods in Ecology and Evolution*, 6(5):557–565. https://doi.org/10.1111/2041-210X.12342.

Miller, D. A., Nichols, J. D., McClintock, B. T., Grant, E. H. C., Bailey, L. L., and Weir, L. A. (2011). Improving occupancy estimation when two types of observational error occur: Non-detection and species misidentification. *Ecology*, 92(7):1422–1428. https://doi.org/10.1890/10-1396.1.

Miller, D. A., Weir, L. A., McClintock, B. T., Grant, E. H. C., Bailey, L. L., and Simons, T. R. (2012). Experimental investigation of false positive errors in auditory species occurrence surveys. *Ecological Applications*, 22(5):1665–1674. https://doi.org/10.1890/11-2129.1.

NEON (2021). Data products: Neon.dp1.10022.001. Provisional data downloaded from https://data.neonscience.org/data-products/DP1.10022.001.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. (2017). How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology letters*, 20(5):561–576. https://doi.org/10.1111/ele.12757.

Plummer, M. et al. (2003). Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Royle, J. A. and Link, W. A. (2006). Generalized site occupancy models allowing for false positive and false negative errors. *Ecology*, 87(4):835–841. https://doi.org/10.1890/0012-9658(2006)87[835:GSOMAF]2.0.CO;2.

Sauer, J. R., Niven, D. K., Hines, J. E., D. J. Ziolkowski, J., Pardieck, K. L., Fallon, J. E., and Link, W. A. (2017). The north american breeding bird survey, results and analysis 1966 - 2015. https://www.mbr-pwrc.usgs.gov/bbs/bbs.html.

Sólymos, P. (2010). dclone: Data cloning in R. *The R Journal*, 2(2):29–37. https://doi.org/10.32614/RJ-2010-011.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A., Lewis, J. S., White, M. D., et al. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, 10(4):585–590. https://doi.org/10.1111/2041-210X.13120.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*. https://arxiv.org/abs/1804.06788.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wright, W. J., Irvine, K. M., Almberg, E. S., and Litt, A. R. (2020). Modelling misclassification in multi-species acoustic data when estimating occupancy and relative activity. *Methods in Ecology and Evolution*, 11(1):71–81. https://doi.org/10.1111/2041-210X.13315.

5 Figures

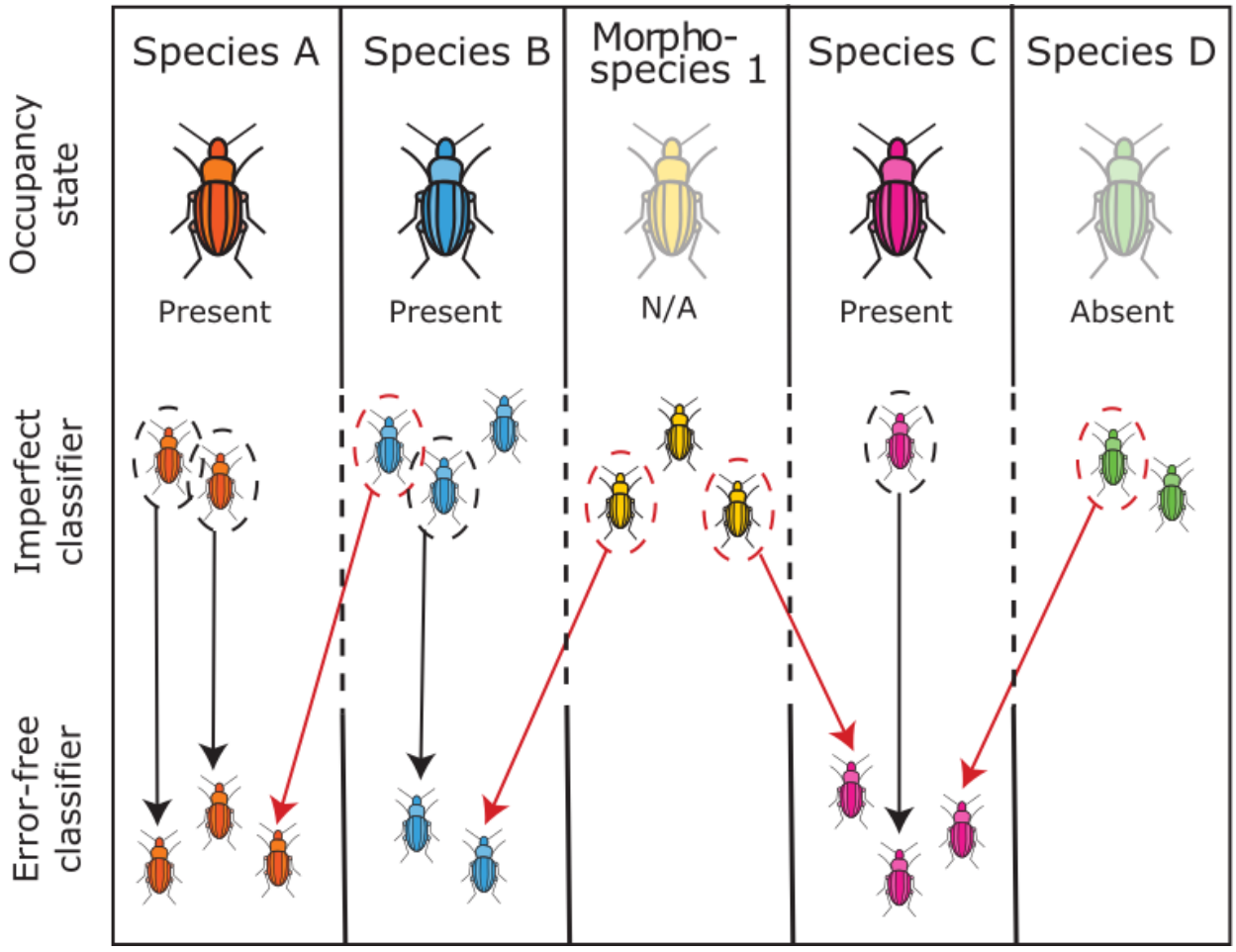

Figure 1: Classification scenarios in NEON carabid data. Each column is an imperfect classifier label. Each species is either present or absent, and morphospecies don't have an occupancy state. In some cases, the imperfect classifier (parataxonomist) matches the error-free classifier (expert taxonomist) (black arrow), in other cases the imperfect classifier was wrong (red arrow), while in other cases still, the error-free classification is unknown due to lack of validation data. For example, the *Morphospecies 1* individual with no error-free classification must belong to a different column, but this species identity is unknown in the raw data.

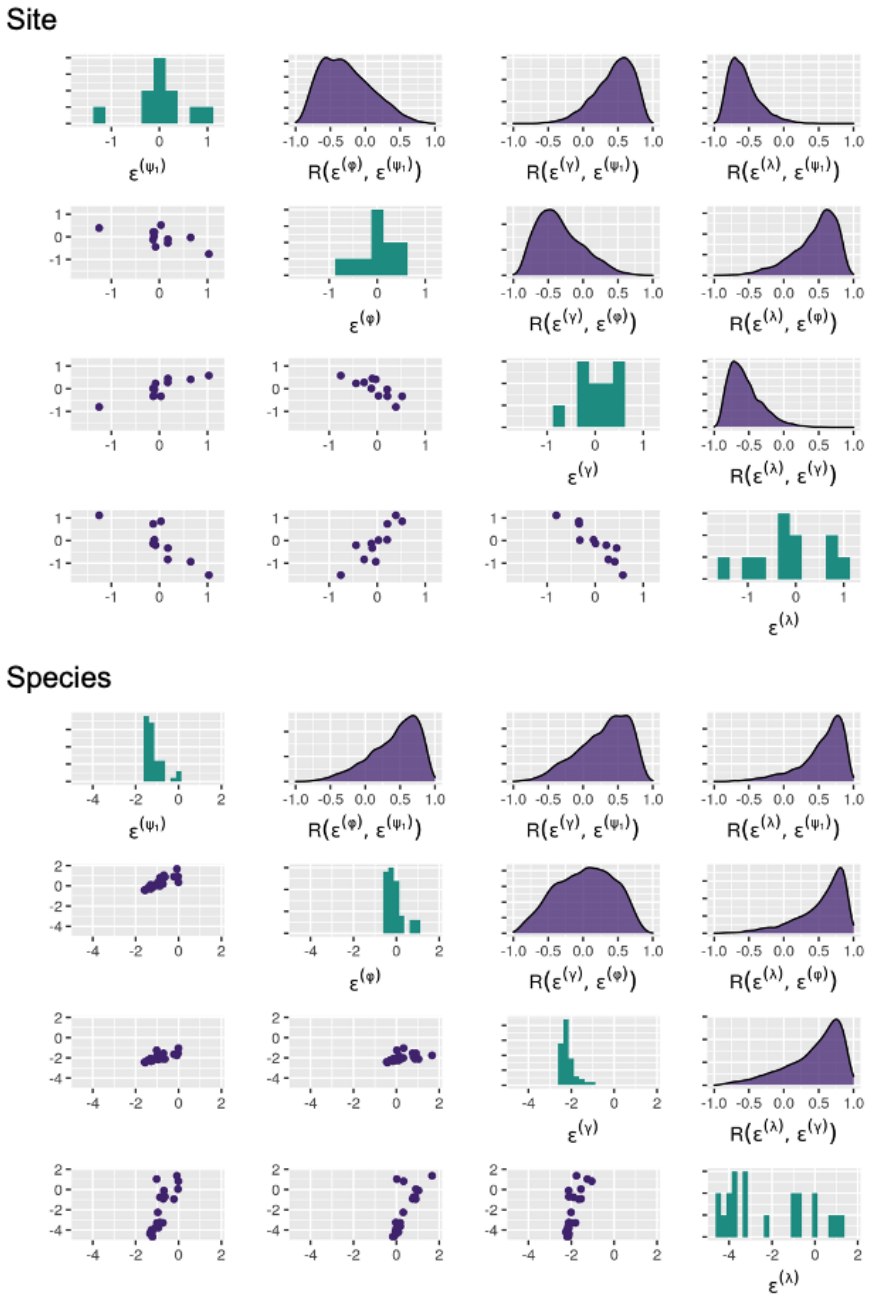

Figure 2: Random effects at the site and species levels. Rows correspond to the rows in the random effect covariance matrix: initial occupancy (row/col 1), persistence (row/col 2), colonization (row/col 3), and encounter rate (row/col 4). Along the diagonal are marginal histograms of posterior medians. Below the diagonal are pairwise scatterplots (each point is a site or species). Above the diagonal are posterior density plots of the pairwise correlation.

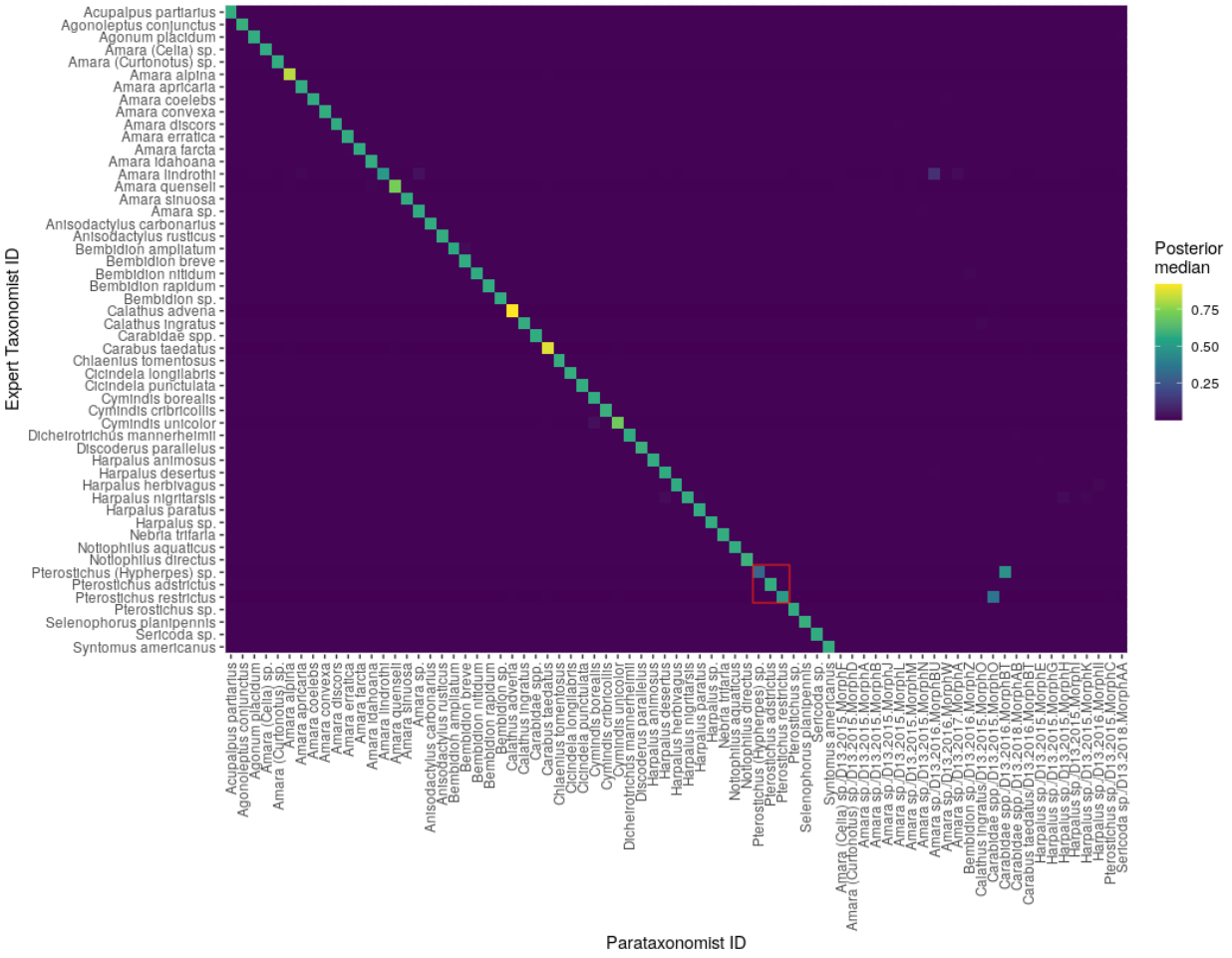

Figure 3: Θ confusion matrix for the full model (classification with occupancy). Heat map of posterior median values where a value is interpreted as the probability that the species in that row is identified as a species in that column. Values along the diagonal are where the species is correctly identified. Values in each row sum to 1. The θ posterior distributions for the 3x3 cells outlined in red are illustrated in Figure 4, along with the posteriors for the reduced (classification alone) and prior models.

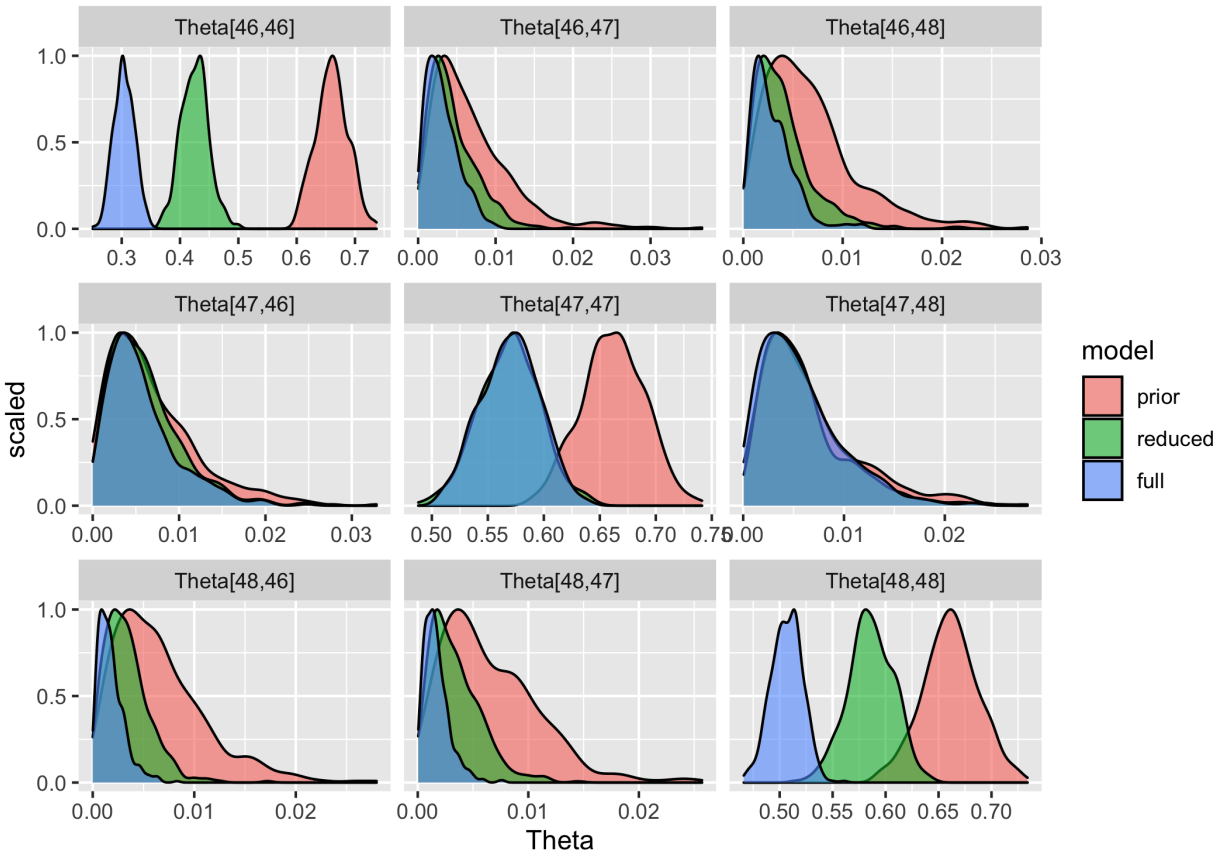

Figure 4: Comparison of θ density distribution between prior, full model posterior (classification with occupancy), and reduced model posterior (classification alone) for select θ confusion matrix indices. For non-abundant species, θ_k probability distributions look like the second row. Here the three models' probability distributions overlap on the off-diagonal (Theta[47,46]/[47,48]) and the posterior distributions overlap but differ from the prior along the diagonal (Theta[47,47]). In contrast the top and bottom rows reflect probability distributions for abundant species. The posterior distributions overlap and are narrower and smaller than the prior on off-diagonal values (Theta[46, 47]/[46,48], Theta[48,46]/[48,47]). Along the diagonal, we see a difference in posterior probability distribution between the reduced and full models (Theta[46,46], Theta[48,48]), visualizing how the full and reduced models perform differently.

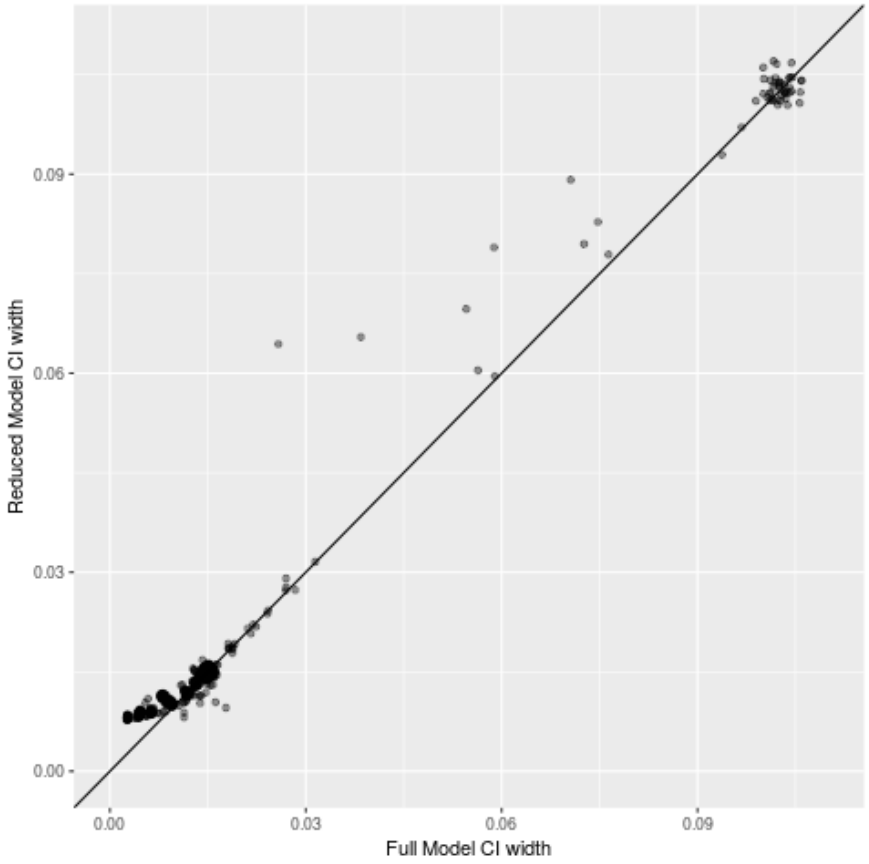

Figure 5: Comparison of precision between theta posterior 95% credible intervals (CI) of full model (occupancy with classification) and reduced model (classification alone). The full model is more precise for points above the line, indicating that the reduced model has a larger CI than the full model for that species.

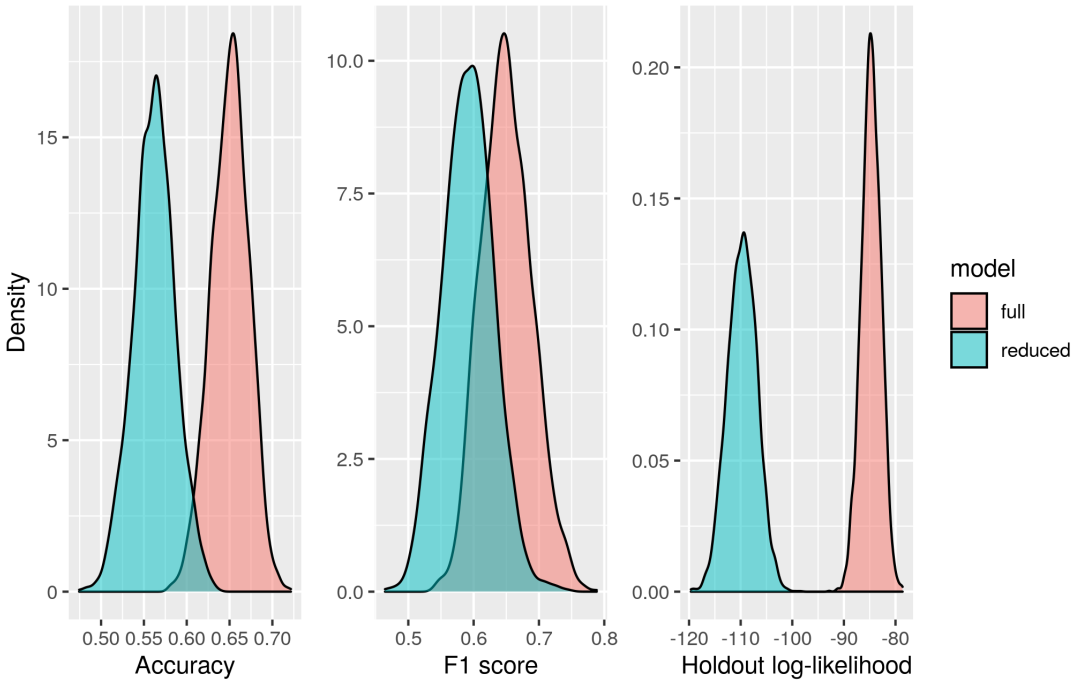

Figure 6: Validation metric macro-averages across posterior draws for accuracy, F1 score, and holdout log-likelihood.

	Full model	Reduced model
Accuracy	0.65	0.56
Precision	0.23	0.19
Recall	0.37	0.35
F1 score	0.65	0.59
Holdout log-likelihood	-84.7	-110

Table 1: Validation metric macro-averages

Appendix S1: multinomial equivalence in the unsupervised setting

The disaggregated dynamic occupancy model that we present is equivalent to the multinomial model of (Wright et al., 2020) in a single-season, unsupervised setting. Fundamentally the equivalence is due to the link between a disaggregated categorical model, and an aggregated multinomial model.

We will outline the equivalence by first considering a set of observations made at one site i and one occasion j, which consist of the total number of individuals encountered $L_{i,j,.}$, and the imperfect species classifications for each encountered individual $y_{i,j,l}$ for $l = 1, ..., L_{i,j,.}$. The disaggregated data model represents the probability of these observations as:

$$[L_{i,j,.} \mid z_{i,1:K}, \lambda_{i,j,1:K}] \prod_{l=1}^{L_{i,j,.}} [y_{i,j,l} \mid k_{i,j,l}, \theta][k_{i,j,l} \mid z_{i,1:K}, \lambda_{i,j,1:K}],$$

where $z_{i,1:K}$ is a vector of binary occupancy states for species $1, ..., K$, $k_{i,j,l}$ is the true species identity of individual l, and $\lambda_{i,j,1:K}$ is a vector of encounter rates.

In this appendix we will show that this is equivalent to the Poisson model of (Wright et al., 2020) (outlined in appendix S1 of their original paper), which consists of conditionally independent Poisson distributions for the number of individuals classified as each species $k' = 1, ..., K$:

$$[C_{i,j,.,1}, ..., C_{i,j,.,K} \mid z_{i,1:K}, \lambda_{i,j,1:K}, \theta] = \prod_{k'=1}^{K} \text{Poisson}\Big(C_{i,j,.,k'} \mid \sum_{k=1}^{K} z_{i,k} \lambda_{i,j,k} \theta_{k,k'}\Big),$$

where $C_{i,j,.,k'}$ is the total number of individuals at site i on occasion j classified as species k', $z_{i,k}$ is the occupancy state of species k, $\lambda_{i,j,k}$ is the expected encounter rate, and $\theta_{k,k'}$ is the probability that an individual of species k is classified as species k'. The proof of equivalence replaces the categorical distribution over class labels in the disaggregated model with a multinomial, then recovers conditionally independent Poisson distributions.

Consider one site and one occasion, dropping the subscripts i and j for simplicity. The model outlined above for the data from this site/occasion is:

$$[L_{.} \mid z_{1:K}, \lambda_{1:K}] \prod_{l=1}^{L_{.}} [y_l \mid k_l, \theta][k_l \mid z_{1:K}, \lambda_{1:K}],$$

where $L_{.}$ is the total number of individuals encountered, y_l is the (imperfect) classification of individual l, and k_l is the true species identity of individual l.

The equivalence holds for the unsupervised case, where none of the true species identities are known. In this case we can marginalize over the unknown species identities as follows:

$$[L_{.} \mid z_{1:K}, \lambda_{1:K}] \prod_{l=1}^{L_{.}} \sum_{k_l=1}^{K} [y_l \mid k_l, \theta][k_l \mid z_{1:K}, \lambda_{1:K}],$$

Writing out the probability mass functions for the Poisson and categorical distributions:

$$= \frac{1}{L_\cdot!}\Big(\sum_{k=1}^{K} z_k\lambda_k\Big)^{L_\cdot} e^{-\sum_{k=1}^{K} z_k\lambda_k} \prod_{l=1}^{L_\cdot}\sum_{k_l=1}^{K} \frac{z_{k_l}\lambda_{k_l}\theta_{k_l,y_l}}{\sum_k z_k\lambda_k}$$

Note that the last term in this expression is the probability mass function for the observed classifications: $[y_l \mid \theta, z_{1:K}, \lambda_{1:K}] = \text{Categorical}\Big(\sum_{k_l=1}^{K} \frac{z_{k_l}\lambda_{k_l}\theta_{k_l,y_l}}{\sum_k z_k\lambda_k}\Big)$. Rewrite this categorical component as a multinomial, where $C_{\cdot,k'}$ represents the total number of individuals classified as species k'.

$$= \frac{1}{L_\cdot!}\Big(\sum_{k=1}^{K} z_k\lambda_k\Big)^{L_\cdot} e^{-\sum_{k=1}^{K} z_k\lambda_k} \frac{L_\cdot!}{\prod_{k'} C_{\cdot,k'}!}\Big(\sum_{k=1}^{K} \frac{z_k\lambda_k\theta_{k,1}}{\sum_k z_k\lambda_k}\Big)^{C_{\cdot,1}} \times ... \times \Big(\sum_{k=1}^{K} \frac{z_k\lambda_k\theta_{k,K}}{\sum_k z_k\lambda_k}\Big)^{C_{\cdot,K}}$$

Note that $L_\cdot!$ cancels out:

$$= \Big(\sum_{k=1}^{K} z_k\lambda_k\Big)^{L_\cdot} e^{-\sum_{k=1}^{K} z_k\lambda_k} \frac{1}{\prod_{k'} C_{\cdot,k'}!}\Big(\sum_{k=1}^{K} \frac{z_k\lambda_k\theta_{k,1}}{\sum_k z_k\lambda_k}\Big)^{C_{\cdot,1}} \times ... \times \Big(\sum_{k=1}^{K} \frac{z_k\lambda_k\theta_{k,K}}{\sum_k z_k\lambda_k}\Big)^{C_{\cdot,K}}$$

Factor out the denominators in the mutinomial probabilities:

$$= \Big(\sum_{k=1}^{K} z_k\lambda_k\Big)^{L_\cdot} e^{-\sum_{k=1}^{K} z_k\lambda_k} \frac{1}{\prod_{k'} C_{\cdot,k'}!}\Big(\frac{1}{\sum_k z_k\lambda_k}\Big)^{\sum_{k'} C_{\cdot,k'}}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,1}} ... \Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,K}}$$

Now, recall that $L_\cdot = \sum_{k'} C_{\cdot,k'}$, which leads to a convenient cancellation:

$$= \frac{\Big(\sum_{k=1}^{K} z_k\lambda_k\Big)^{L_\cdot}}{\Big(\sum_{k=1}^{K} z_k\lambda_k\Big)^{L_\cdot}} e^{-\sum_{k=1}^{K} z_k\lambda_k} \frac{1}{\prod_{k'} C_{\cdot,k'}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,1}} ... \Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,K}}$$

$$= e^{-\sum_{k=1}^{K} z_k\lambda_k} \frac{1}{\prod_{k'} C_{\cdot,k'}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,1}} ... \Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,K}}$$

Group like terms:

$$= e^{-z_1\lambda_1} ... e^{-z_K\lambda_K} \frac{1}{C_{\cdot,1}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,1}} ... \frac{1}{C_{\cdot,K}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,K}}$$

Then, note that $\exp(-z_k\lambda_k) = \exp(-\sum_{k'=1}^{K} z_k\lambda_k\theta_{k,k'})$:

$$= \prod_{k=1}^{K} e^{-\sum_{k'=1}^{K} z_k\lambda_k\theta_{k,k'}} \frac{1}{C_{\cdot,1}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,1}} ... \frac{1}{C_{\cdot,K}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,K}}$$

$$= \prod_{k=1}^{K}\prod_{k'=1}^{K} e^{-z_k\lambda_k\theta_{k,k'}} \frac{1}{C_{\cdot,1}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,1}} ... \frac{1}{C_{\cdot,K}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,K}}$$

Finally, we can group these terms to recover the conditionally independent Poisson random variables from (Wright et al., 2020):

$$= \left(\frac{1}{C_{\cdot,1}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}\Big)^{C_{\cdot,1}} e^{-\sum_{k=1}^{K} z_k\lambda_k\theta_{k,1}}\right) ... \left(\frac{1}{C_{\cdot,K}!}\Big(\sum_{k=1}^{K} z_k\lambda_k\theta_{k,K}\Big)^{C_{\cdot,K}} e^{-\sum_{k=1}^{K} z_k\lambda_k\theta_{k,K}}\right)$$

$$= \prod_{k'=1}^{K} \text{Poisson}\Big(C_{.,k'} \mid \sum_{k=1}^{K} z_k \lambda_k \theta_{k,k'}\Big),$$

which proves the equivalence between the disaggregated categorical observation model and the observation model described in (Wright et al., 2020).

Appendix S2: case study model details

The NEON carabid beetle case study estimates occupancy dynamics and encounter rates for sites $i = 1, ..., N$, species $k = 1, ..., K$, and years $t = 1, ..., T$. The state and observation models are identical to those described in the main text. Occupancy dynamics were modeled as a first order Markov process. For the first year:

$$z_{i,k1} \sim \text{Bernoulli}(\psi_{i,k,1}).$$

In subsequent years:

$$z_{i,k,t} \sim \text{Bernoulli}(z_{i,k,t-1}\phi_{i,k,t-1} + (1 - z_{i,k,t-1})\gamma_{i,k,t-1}),$$

where $z_{i,k,t}$ is a binary occupancy state, $\phi_{i,k,t}$ is the probability of a site remaining occupied from year t to year $t+1$, and $\gamma_{i,k,t}$ is the probability of an unoccupied site in year t being occupied in year $t+1$.

We assume that encounter rates vary by site and species, leading to the following encounter model for occasion j:

$$L_{i,j,.,t} \sim \text{Poisson}\Big(\sum_{k=1}^{K} z_{i,k,t}\lambda_{i,k}\Big).$$

The model for the observed species or morphospecies classifications is:

$$y_{i,j,l,t} \sim \text{Categorical}(\boldsymbol{\theta}_{k[i,j,l,t]}),$$

Classification probabilities were modeled using an informative Dirichlet prior:

$$\boldsymbol{\theta}_k \sim \text{Dirichlet}(\boldsymbol{\alpha}_k),$$

where $\alpha_{k,k}$ was set to 80, and all other elements $\alpha_{k,k'}$ such that $k \neq k'$ were set to 2. These specific values were chosen based on prior predictive simulations to match our expectations about the classification skill of parataxonomists.

The model for true species identities is:

$$k[i,j,l,t] \sim \text{Categorical}\left(\frac{z_{i,k,t}\lambda_{i,k}}{\sum_k z_{i,k,t}\lambda_{i,k}}\right).$$

We modeled initial occupancy ($\psi_{i,k,1}$), persistence ($\phi_{i,k}$), colonization ($\gamma_{i,k}$), and encounter rates ($\lambda_{i,k}$) as a function of site and species-specific random effects:

$$\text{logit}(\psi_{i,k,1}) = \epsilon_{i,1} + \alpha_{k,1},$$

$$\text{logit}(\phi_{i,k}) = \epsilon_{i,2} + \alpha_{k,2},$$

$$\text{logit}(\gamma_{i,k}) = \epsilon_{i,3} + \alpha_{k,3},$$

$$\log(\lambda_{i,k}) = \epsilon_{i,4} + \alpha_{k,4}.$$

These random effects were modeled with multivariate normal priors to allow for correlations among parameters. At the site level:

$$\begin{bmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \epsilon_{i,3} \\ \epsilon_{i,4} \end{bmatrix} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^{(\epsilon)}).$$

At the species level:

$$\begin{bmatrix} \alpha_{k,1} \\ \alpha_{k,2} \\ \alpha_{k,3} \\ \alpha_{k,4} \end{bmatrix} \sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma}^{(\alpha)}).$$

The covariance matrices associated with these random effects were given inverse Wishart priors, with the parameters of the inverse Wishart distributions chosen based on prior predictive simulations:

$$\boldsymbol{\Sigma}^{(\epsilon)} \sim \text{Wishart}^{-1}(\boldsymbol{I}, 10),$$

$$\boldsymbol{\Sigma}^{(\alpha)} \sim \text{Wishart}^{-1}(\boldsymbol{I}, 10),$$

where $\boldsymbol{I}$ is a 4×4 identity matrix, and 10 represents the degrees of freedom parameter of the inverse Wishart distribution. Finally the vector of means, which acts essentially as an intercept parameter, was given a standard normal prior: $\boldsymbol{\mu} \sim \text{Normal}(\mathbf{0}, \boldsymbol{I})$. A graphical representation of the model is given in Figure S2.1.

We used simulation-based calibration (SBC) to check the validity of the sampler implemented by JAGS (Talts et al., 2018). Simulation-based calibration involves simulating many datasets from the prior predictive distribution of the model, and for each simulated dataset, drawing samples from the posterior. These posterior samples can then be compared to the true (known) values simulated from the prior in terms of rank statistics, which for well-behaved samplers are uniform. Non-uniform rank statistics are indicative of problems with the sampler, e.g., a tendency to under- or over-sample from the tails, or systematic directional bias (Talts et al., 2018). We simulated 3000 data sets from the prior distribution of the model in the two-species case, with 20 sites, 5 time steps, and 2 sampling occasions per site in each time step. Uniform SBC rank histograms indicate that the MCMC sampler built by JAGS is able to draw samples from the posterior distribution of the model (Fig. S2.2).

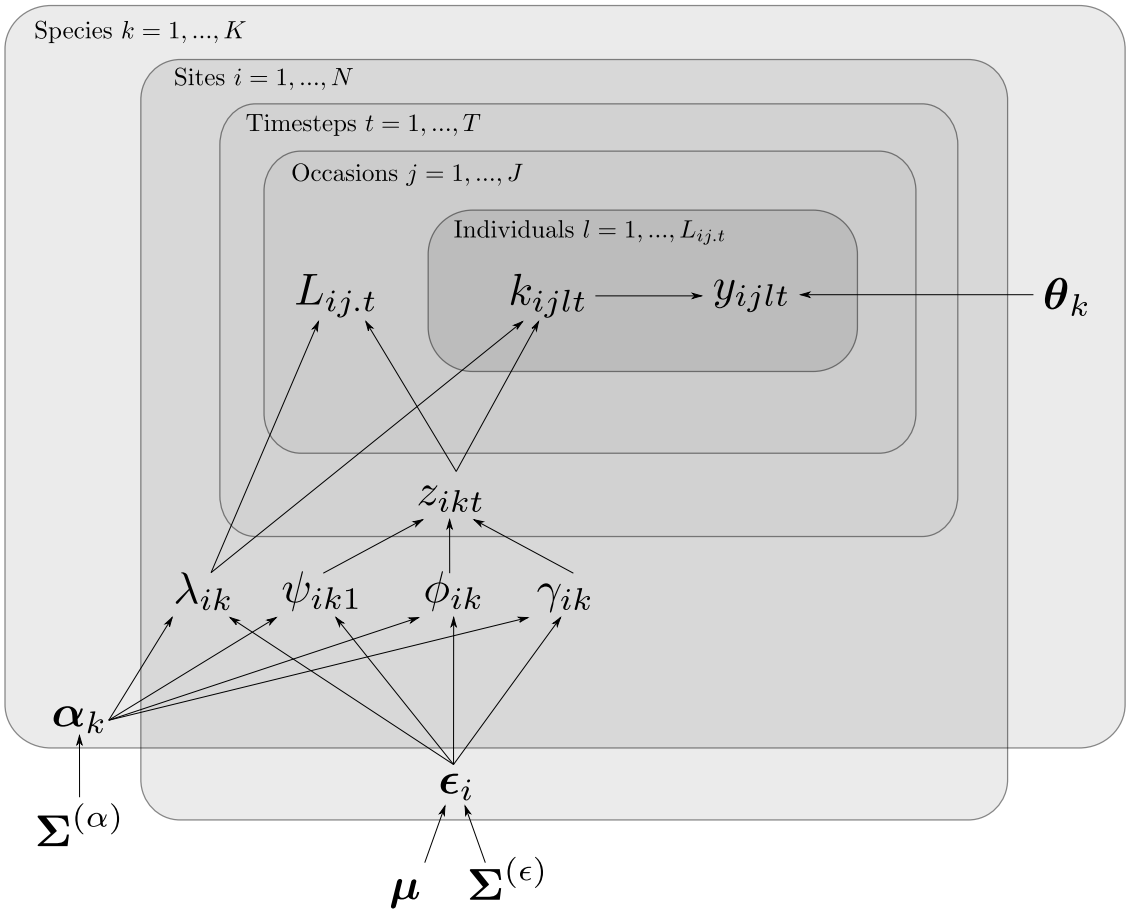

Figure S2.1: Directed acyclic graph representation of the NEON carabid beetle model. Grey plates represent levels of the model and edges represent dependence.

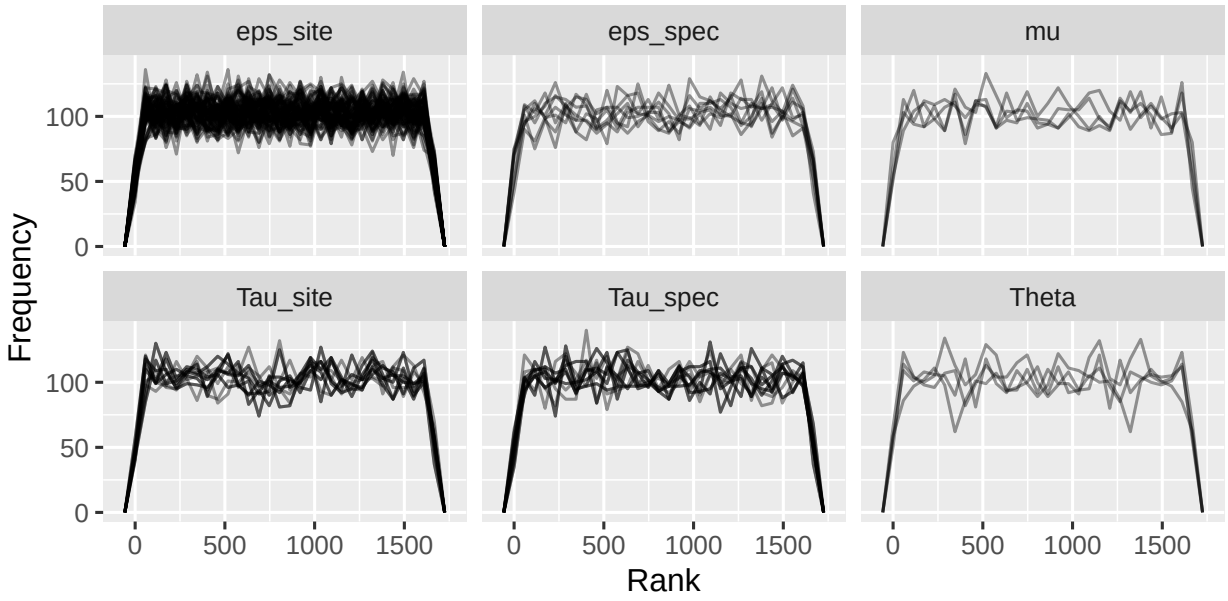

Figure S2.2: Simulation based calibration rank frequency polygons indicate uniformity. Each panel represents a different model parameter group, and each line represents a different parameter in that group (e.g., there are two species, so that the matrix $\mathbf{\Theta}$ is 2 by 2 so that there are four elements, and there are four lines in the Theta panel, one for each matrix element).